Note on Probability and Statistics (Undergraduate)

Zepeng CHEN
The HK PolyU

Date: March 8, 2023

1 Probability

1.1 Basics of Probability Theory

Definition 1.1 (Sample space, event)

- *The set, Ω, of all possible outcomes of a particular experiment is called the sample space for the experiment.*
- *An event is any collection of possible outcomes of an experiment, that is, any subset of Ω (including Ω itself).*

Theorem 1.1

- $P\left(B \cap A^{c}\right)=P(B)-P(A \cap B)$
- $P(A \cup B)=P(A)+P(B)-P(A \cap B)$
- *Multiplication rule:*

$$\mathbf{P}(A \cap B \cap C)=\mathbf{P}(A) \cdot \mathbf{P}(B \mid A) \cdot \mathbf{P}(C \mid A \cap B)$$

Corollary 1.1 (Boole's Inequality)

$P\left(\cup_{i=1}^{\infty} A_{i}\right) \leq \sum_{i=1}^{\infty} P\left(A_{i}\right)$ *for any sets* A_1, A_2, ...

Corollary 1.2 (Bonferroni's Inequality)

$$P(A \cap B) \geq P(A)+P(B)-1 \quad P\left(\bigcap_{i=1}^{n} A_{i}\right) \geq \sum_{i=1}^{n} P\left(A_{i}\right)-(n-1)$$

Note on *Bonferroni's Inequality allows us to bound the probability of a simultaneous event (the intersection) in terms of the probabilities of the individual events. Note that unless the probabilities of the individual events are sufficiently large, the Bonferroni bound is a useless (but correct!) negative number. For example, suppose A and B are two events and each has probability $.95$, then*

$$P(A \cap B) \geq P(A)+P(B)-1=.95+.95-1=.90.$$

	Without replacement	With replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	C_n^r	C_{n+r-1}^r

1.2 Counting

Definition 1.2 (Factorial)

$n!$ (n factorial) is defined by

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1.$$

Note on *Furthermore, we define $0! = 1$.*

Note on Stirling's Formula

Theorem 1.2 (Fundamental Theorem of Counting)

If a job consists of k seperate tasks, the ith of which can be done in n_i ways, $i = 1, ..., k$, then the entire job can be done in $n_1 \times n_2 \times \cdots \times n_k$ ways.

Note on *The distinction between counting with replacement and counting without replacement, and the ordering of the tasks are important. For example, the New York state lottery operated according to the following scheme. From the numbers 1, 2, ...,44, a person may pick any six for her ticket. There are four cases.*

- ***Ordered, without replacement.*** $44 \times 43 \times 42 \times 41 \times 40 \times 39 = \frac{44!}{38!}$.
- ***Ordered, with replacement.*** $44 \times 44 \times 44 \times 44 \times 44 \times 44 = 44^6$.
- ***Unordered, without replacement.*** *We must divide out the redundant orderings, note that six numbers can be arranged in $6 \times 5 \times 4 \times 3 \times 2 \times 1$ ways, so*
$$\frac{44 \times 43 \times 42 \times 41 \times 40 \times 39}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{44!}{6!38!}.$$
- ***Unordered, with replacement.*** *The first guess that $44^6/(6 \times 5 \times 4 \times 3 \times 2 \times 1)$ is not correct. To count in this case, it is easiest to think of placing 6 markers on the 44 numbers. In fact, we can think of the 44 numbers defining bins in which we can place the six markers, M. Thus, we have to count all of the arrangements of 43 walls (44 bins yield 45 walls, but we disregar the two end walls) and 6 markers. We therefore have 43+6=49 objects,which can be arranged in 49! ways. However, to eliminate the redundant orderings we must divide by both 6! and 43!, so $\frac{49!}{6!43!}$.*

To summarize, the number of possible arrangements of size r from n objects are

1.3 Conditional Probability and Independence

Definition 1.3 (Conditional probability)

The conditional prob of A given B, is $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$.

Note on Example *An interesting example shows that conditional probabilities require careful interpretation.*

Theorem 1.3

- *Total probability theorem:*

$$\mathrm{P}(B)=\mathrm{P}\left(A_1\right)\mathrm{P}\left(B \mid A_1\right)+\mathrm{P}\left(A_2\right)\mathrm{P}\left(B \mid A_2\right)+\mathrm{P}\left(A_3\right)\mathrm{P}\left(B \mid A_3\right)$$

- *Bayes' rule:*

$$\begin{aligned}\mathrm{P}\left(A_i \mid B\right) &= \frac{\mathrm{P}\left(A_i \cap B\right)}{\mathrm{P}(B)} \\ &= \frac{\mathrm{P}\left(A_i\right)\mathrm{P}\left(B \mid A_i\right)}{\mathrm{P}(B)} \\ &= \frac{\mathrm{P}\left(A_i\right)\mathrm{P}\left(B \mid A_i\right)}{\sum_j \mathrm{P}\left(A_j\right)\mathrm{P}\left(B \mid A_j\right)}\end{aligned}$$

Definition 1.4 (Independence)

- *Two events E and F are independent if $P(E\cap F)=P(E)P(F)$, the event $E_1...E_n$ are independent if for every subset of these events $P(E_{i_1}...E_{i_r})=P(E_{i_1})...P(E_{i_r})$.*
- *The random variables X and Y are independent if $F(x,y)=F_X(x)F_Y(y)$, the n random variables are independent if $F(x_1,..x_n)=F_{X_1}(x_1)...F_{X_n}(x_n)$.*

Note on *A pair of events A and B cannot be simultaneously mutually exclusive and independent. Otherwise, if $P(A)>0$ and $P(B)>0$, then $P(AB)=0=P(A)P(B)$, which contradicts our assumption.*

Theorem 1.4

If A and B are independent events, then the following pairs are also independent:

- *A and B^c,*
- *A^c and B,*
- *A^c and B^c.*

1.4 Random Variables

Definition 1.5

A random variable is a function from a sample space Ω into the real numbers.

Note on *R.V. can be used to reduce the size of the problem. For example, suppose we collect "yes" or "no" in an opinion poll, then the sample space has 2^50 elements. If we define a variable X as the number of "yes", then the sample space is the set of integers $\{0,...,50\}$.*

1.5 Distribution Functions

Definition 1.6 (CDF)

- *The cumulative distribution function is defined by*

$$F_X(x) = P_X(X \leq x), \textit{ for all } x.$$

Note on *F_X can be discontinuous, with jumps at certain values of x. However, at the jump points F_X takes the value at the top of the jump. That is, right-continuity. The property of right-continuity is a consequence of the definition of the cdf.*

Note on Complementary distribution function *Complementary distribution function is defined by $\bar{F}(x) = 1 - F(x) = P\{X > x\}$.*

Definition 1.7

A random variable X is continuous if $F_X(x)$ is a continuous function of x. A random variable X is discrete if $F_X(x)$ is a step function of x.

Definition 1.8

The random variables X and Y are identically distributed if, for every set $A \in \mathcal{B}^1$, $P(X \in A) = P(Y \in A)$.

Note on *Note that this does not say that $X = Y$.*

Theorem 1.5

The following two statements are equivalent:

- *The random variables X and Y are identically distributed.*
- *$F_X(x) = F_Y(y)$ for every x.*

1.6 Density and Mass Functions

Definition 1.9 (PMF)

The probability mass function is given by $f_X(x) = P(X = x)$ for all x.

Definition 1.10 (PDF)

The probability density function is given by

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \quad \textit{for all } x.$$

1.7 Moment Generating Function

Definition 1.11 (Moment Generating Function)

For a random variable X, the moment generating function $M_X : \mathbb{R} \to \mathbb{R}$ is defined as

$$M_X(t) := \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} p_X(x)e^{tx}dx.$$

Lemma 1.1

If X_1 and X_2 are independent random variables, then the MGF of their sum $X_1 + X_2$ is the product of their MGFs:

$$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t).$$

Lemma 1.2 (MGF of Gaussian Variables)

Suppose $X \sim \mathcal{N}(0, \sigma^2)$ is a zero-mean Gaussian variable with variance σ^2. Then

$$M_X(t) = \exp(\frac{\sigma^2 t^2}{2}).$$

Proof

$$\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2}) \exp(tx)dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-t\sigma^2)^2}{2\sigma^2} + \frac{t^2\sigma^2}{2})dx \quad \text{(Completing the square)} \\
&= \exp(\frac{t^2\sigma^2}{2}) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-t\sigma^2)^2}{2\sigma^2})dx = \exp(\frac{t^2\sigma^2}{2}) \quad \text{(Normal dist.)}
\end{aligned}$$

■

2 Discrete distributions

2.1 Geometric distribution

3 Continuous distributions

3.1 TBD

Theorem 3.1 (Jensen's Inequality)

For any concave function f, we have

$$E[f(X)] \leq f(E(X))$$

Particularly, if $f(x) = \frac{1}{x}$, we have

$$E[X]E[\frac{1}{X}] \leq 1$$

Similarly, for any convex function f, we have

$$E[f(X)] \geq f(E(X))$$

Theorem 3.2 (Chebyshev's Inequality)

Let X be a random variable with $E[X]$ and $Var[X]$, then

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2} \quad \textit{or} \quad P(|X| \geq a) \leq \frac{EX^2}{a^2}$$

Proof

$$I\{|X| \geq a\} \leq X^2/a^2$$

Then the result follows by taking expectations. ■

Theorem 3.3 (Markov's Inequality)

If X is a nonnegative random variable, then for any $a > 0$

$$P\{X \geq a\} \leq E[X]/a$$

Proof Let $I\{X \geq a\}$ be 1 if $X \geq a$ and 0 otherwise. Then it is easy to see since $X \geq 0$ that

$$aI\{X \geq a\} \leq X$$

Taking expectations yields the result. ■

Theorem 3.4 (Gibbs' Inequality (Soch et al., 2022, p. 94))

Let X be a discrete random variable and consider two probability distributions with pmf $p(x)$ and $q(x)$. Then the entropy of X according to P is smaller than or equal to the cross-entropy of P and Q:

$$-\sum_{x \in \mathcal{X}} p(x) log_b p(x) \leq -\sum_{x \in \mathcal{X}} p(x) log_b q(x).$$

Proof This is equivalent to show the KL-divergence is non-negative, i.e.,

$$\sum_{x \in \mathcal{X}} p(x) log_b \frac{p(x)}{q(x)} \geq 0.$$

Next, we can prove it via showing that $f(x) = lnx - (x-1) \leq 0$ if $x > 0$ by its concavity. ■

4 Statistics

4.1 Expectation

Definition 4.1 (Expectation)

$$\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} x dF(x) = \begin{cases} \int_{-\infty}^{\infty} x f(x) dx & \textit{if } X \textit{ is continuous} \\ \sum_x x P\{X = x\} & \textit{if } X \textit{ is discrete} \end{cases} \\
E[h(X)] &= \int_{-\infty}^{\infty} h(x) dF(x) \quad \textit{if } X \textit{ is continuous} \\
E[\sum_{i=1}^{n} X_i] &= \sum_{i=1}^{n} E[X_i] \quad E[aX] = aE[X]
\end{aligned} \tag{1}$$

Lemma 4.1 (Tonelli's theorem)

If $x_{ki} \geq 0$ for all k, i, then

$$\sum_{k=1}^{\infty} \sum_{i=1}^{k} x_{ki} = \sum_{i=1}^{\infty} \sum_{k=i}^{\infty} x_{ki}$$

Remark Note that when x is negative, the equation may not remain.

Lemma 4.2 (Expectation of nonnegative integer-valued random variable)

$$E[N] = \sum_{i=1}^{\infty} P\{N \geq i\} = \sum_{i=0}^{\infty} P\{N > i\}$$

Proof

$$\begin{aligned}
E[N] &= \sum_{k=1}^{\infty} k P\{N = k\} \\
&= \sum_{k=1}^{\infty} \sum_{i=1}^{k} P\{N = k\} \qquad (k = \sum_{i=1}^{k} 1) \\
&= \sum_{i=1}^{\infty} \sum_{k=i}^{\infty} P\{N = k\} \quad (\text{Theorem } 4.1) \\
&= \sum_{i=1}^{\infty} P\{N \geq i\}
\end{aligned} \tag{2}$$

■

Lemma 4.3 (Expectation of Nonnegative Random Variables)

For any nonnegative random variable X,

$$\begin{aligned}
E[X] &= \int_{0}^{\infty} \bar{F}(x) dx \\
E[X^n] &= \int_{0}^{\infty} n x^{n-1} \bar{F}(x) dx
\end{aligned} \tag{3}$$

Remark This lemma is the generalization of Lemma 4.2

Proof

$$
\begin{aligned}
E\left[X^n\right] &= \int_0^{\infty} x^n dF(x) \\
&= \int_0^{\infty} \int_0^{x} n t^{n-1} dt dF(x) \\
&= \int_0^{\infty} \int_t^{\infty} n t^{n-1} dF(x) dt \\
&= \int_0^{\infty} n t^{n-1} \cdot [F(\infty) - F(t)] dt \\
&= \int_0^{\infty} n t^{n-1} \bar{F}(t) dt
\end{aligned}
$$

■

4.2 Variance

Definition 4.2 (Variance)

$$
\begin{aligned}
Var(X) &= E[(X - E[X])^2] = E[X^2] - E^2[X] \\
Var[\sum_{i=1}^{n} X_i] &= \sum_{i=1}^{n} Var(X_i) + 2\sum\sum_{i<j} Cov(X_i, X_j) \\
\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \operatorname{Var}\left(X_i\right) \quad \textit{Independent} \\
\operatorname{Var}(aX) &= a^2 \operatorname{Var}(X)
\end{aligned}
\tag{4}
$$

Definition 4.3 (Covariance)

$$
Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y] \tag{5}
$$

Lemma 4.4 (Covariance inequality)

$$
|\operatorname{Cov}(X, Y)| \leq \sigma_X \sigma_Y \ \textit{or} \ \operatorname{Cov}^2(X, Y) \leq \operatorname{Var}(X) \operatorname{Var}(Y)
$$

Proof This inequality can be easily proved by Cauchy–Schwarz inequality. ■

Definition 4.4 (correlation coefficient)

Note that $-1 \leq \rho \leq 1$, and $|\rho| = 1$ means that X and Y are linearly related, independent means that $|\rho| = 0$, but the converse is not true.

$$
\begin{aligned}
\rho &= \mathrm{E}\left[\frac{(X - \mathrm{E}[X])}{\sigma_X} \cdot \frac{(Y - \mathrm{E}[Y])}{\sigma_Y}\right] \\
&= \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}
\end{aligned}
$$

Remark X and Y are uncorrelated if $\operatorname{Cov}(X, Y) = 0$. Independent are uncorrelated. However, the converse needs not be true.

Definition 4.5 (Coefficient of variation)
standard deviation divided by mean.

Remark It is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.

Example 4.1Bounded variable's variance must be bounded If $P\{0 \le X \le a\} = 1$, show that

$$Var[X] \le \frac{a^2}{4}$$

Proof [First Proof] Define $Y = X - \frac{a}{2}$, easy to know that $Var[X] = Var[Y]$, thus our goal turns to prove $Var[Y] \le \frac{a^2}{4}$

Since $0 \le X \le a$, we have $-\frac{a}{2} \le Y \le \frac{a}{2}$, thus $E[Y^2] \le \frac{a^2}{4}$ ($E[Y^2] = \int_0^{\frac{a^2}{4}} s f_{Y^2}(s) ds \le \frac{a^2}{4} \int_0^{\frac{a^2}{4}} f_{Y^2}(s) ds = \frac{a^2}{4}$).

$$\operatorname{Var}(Y) = E\left[Y^2\right] - (E[Y])^2 \le \frac{a^2}{4} - (E[Y])^2 \le \frac{a^2}{4}$$

■

Proof [Another Proof] Firstly, we have $E[X^2] \le E[aX] = aE[X]$. Note that

$$\operatorname{Var}(X) = E\left[X^2\right] - E[X]^2 \le a\left(E[X] - \frac{E[X]^2}{a}\right)$$

$$\begin{aligned} &(2E[X] - a)^2 \\ =&4E[x]^2 - 4aE[x] + a^2 \\ =&4a\left(\frac{E[X]^2}{a} - E[x] + \frac{a}{4}\right) \\ \ge&0 \end{aligned}$$

Hence, we have the follows, then we prove it.

$$E[X] - \frac{E[X]^2}{a} \le \frac{a}{4}$$

■

Bibliography

Soch, Joram et al. (Jan. 2022). *StatProofBook/StatProofBook.Github.Io: StatProofBook 2021*. Zenodo. DOI: 10.5281/zenodo.5820411.